Thinning and Information Projections

Peter Harremoës
Stat. learning and quant. comp.
Centrum voor Wiskunde en Informatica
P.O. 94079, 1090 GB Amsterdam
The Netherlands
P.Harremoes@cwi.nl

Oliver Johnson
Dept. Mathematics
University of Bristol
Bristol, BS8 1TW
United Kingdom
O.Johnson@bristol.ac.uk

Ioannis Kontoyiannis
Athens University
of Econ. and Business
Patission 76, Athens 10434
Greece
yiannis@aueb.gr

***Abstract*— The law of thin numbers is a Poisson approximation theorem related to the thinning operation. We use information projections to derive lower bounds on the information divergence from a thinned distribution to a Poisson distribution. Conditions for the existence of projections are given. If an information projection exists it must be an element of the associated exponential family. Exponential families are used to derive lower bounds on information divergence and lower bounds on the rate of convergence in the law of thin numbers. A method of translating results related to Poisson distributions into results related to Gaussian distributions is developed and used to prove a new non-trivial result related to the central limit theorem.**

I. Introduction

Approximation by a Poisson distribution is a well studied subject and a careful presentation can be found in [1]. Connections to information theory have been established in [2], [3]. For most values of the parameters, the best bounds on total variation between a binomial distribution and a Poisson distribution with the same mean have been proved by ideas from information theory via Pinsker's inequality [4], [5], [6], [7]. Recently [8] the idea of thinning a random variable was introduced and used to formulate and prove a Law of Thin Numbers that is a way of formulating the Law of Small Numbers (Poisson's Law) so that it resembles formulation of the Central Limit Theorem for a sequence of independent identically distributed random variables. Here these ideas will be developed further. There are three main reasons for developing these results. The first is to get a lower bound for the rate of convergence in the Law of Thin Numbers, the second is to use these to get new inequalities and asymptotic results for the central limit theorem, and the last is to develop the general understanding and techniques related to information divergence and information projection. Many of our calculations involve Poisson-Charlier polynomials and are quite lengthy. Many details have been left out and short versions of the proofs can be found in the appendix. We hope eventually to be able to tell which aspects of important theorems for continuous variables like the Entropy Power Inequality that can be derived from results for discrete variables and which aspect are essentially related to continuous variables. The relevance for communication will not be discussed here, see [8] for some related results.

II. Preliminaries on Thinning

Let P denote a distribution on $\mathbb{N}_0$. For $\alpha \in [0;1]$ the α-thinning of P is the distribution $T_\alpha(P)$ given by

$$T_\alpha(P)(k) = \sum_{l=k}^{\infty} P(l) \binom{l}{k} \alpha^k (1-\alpha)^{l-k}.$$

If $X_1, X_2, X_3, \ldots$ are independent identically distributed Bernoulli random variables with success probability α and Y has distribution P independent of $X_1, X_2, \cdots$ then

$$\sum_{n=1}^{Y} X_n$$

has distribution $T_\alpha(P)$. Obviously the thinning of an independent sum of random variables is the convolution of thinnings. We shall use the notation $x^{\underline{k}} = x(x-1)\cdots(x-k+1)$. The factorial moments of an α-thinning are easy to calculate

$$E\left[\left(\sum_{n=1}^{Y} X_n\right)^{\underline{k}}\right] = E\left[E\left[\left(\sum_{n=1}^{Y} X_n\right)^{\underline{k}} \middle| Y\right]\right] = E\left[\alpha^k Y^{\underline{k}}\right] = \alpha^k E\left[Y^{\underline{k}}\right]. \quad (1)$$

Thus, thinning scales the factorial moments in the same way as ordinary multiplication scales the ordinary moments.

Thinning transforms binomial distributions into binomial distributions, Poisson distributions into Poisson distributions, geometric distributions into geometric distributions and negative binomial distributions into negative binomial distributions [8]. A distribution on $\mathbb{N}_0$ is said to be ultra log-concave if its density with respect to a Poisson distribution is discrete log-concave. Thinning also conserves the class of ultra log-concave distributions [9].

The thinning operation allow us to state and prove the Law of Thin Numbers in various versions [8].

Theorem 1: Let P be a distribution on $\mathbb{N}_0$ with mean λ. Then $T_{1/n}(P^{*n})$ converges pointwise to $\mathrm{Po}(\lambda)$ as $n \to \infty$. If P is an ultra log-concave distribution on $\mathbb{N}_0$ then

$$H\left(T_{1/n}(P^{*n})\right) \to H(\mathrm{Po}(\lambda)), \quad \text{as } n \to \infty,$$

where P^{*n} denote the n-fold convolution of P. If the divergence $D(P\|\mathrm{Po}(\lambda))$ is finite then

$$D\left(T_{1/n}(P^{*n})\|\mathrm{Po}(\lambda)\right) \to 0, \quad \text{as } n \to \infty.$$

The main aim of this paper is to develop techniques that allow us to give lower bounds on the rate of convergence in the Law of Thin Numbers.

III. Existence of minimum information distributions

Let X be a random variable for which the moments of order $1, 2, ...l$ exist. We shall assume that $E(X) = \lambda$. We are interested in minimizing information divergence $D(X\|\mathrm{Po}(\lambda))$ under linear conditions on the moments of X and derive conditions for a minimum divergence distribution to exist. For the results we shall derive later it will be convenient that we calculate moments with respect to the Poisson-Charlier polynomials, which are given by the following definition.

Definition 2: The *Poisson-Charlier polynomial* of order k is given by

$$P_k(x) = \left(\lambda^k k!\right)^{-1/2} \sum_{l=0}^{k} \binom{k}{l} (-\lambda)^{k-l} x^{\underline{l}}.$$

The Poisson-Charlier polynomials are characterized as normalized orthogonal polynomials with respect to the Poisson distribution $\mathrm{Po}(\lambda)$.

Lemma 3: For some fixed set $(h_1, \cdots, h_l) \in \mathbb{R}^l$, let K be the convex set of distributions on $\mathbb{N}_0$ for which the first l moments are defined and which satisfies the following conditions

$$E_P[P_k(X)] = h_k \ , \text{ for } k = 1, 2, \cdots, l-1; \tag{2a}$$
$$E_P[P_l(X)] \le h_l \ . \tag{2b}$$

If $K \neq \varnothing$ then the minimum information projection of $\mathrm{Po}(\lambda)$ exists.

Theorem 4: Let C be the set of distributions on $\mathbb{N}_0$ for which the first l moments are defined and satisfy the following equations

$$E[P_k(X)] = h_k \text{ for } k = 1, 2, \cdots, l. \tag{3}$$

Assume that $C \neq \varnothing$ and $l \geq 2$. We shall consider the following three cases:

1) $h_k = 0$ for $k < l$ and $h_l > 0$.
2) $h_k = 0$ for $k < l$ and $h_l < 0$.
3) $h_k = 0$ for $k < l-1$ and $h_{l-1} > 0$.

In case 1 no minimizer exists and $\inf_C D(P\|Po(\lambda)) = 0$. In case 2 and 3 there exists a distribution P^* in C that minimizes $D(P\|Po(\lambda))$.

Proofs can be found in the appendix. For our applications it is easy to check that the set C defined in Theorem 4 is non-empty, but in general it may be difficult to determine simple necessary and sufficient conditions for $C \neq \varnothing$ in terms of the set of specified moments.

IV. Lower bounds

Let X be a random variable with values in $\mathbb{N}_0$ and with mean μ. Then the divergence $D(X\|\mathrm{Po}(\lambda))$ is minimal if the distribution of X is element of the associated exponential family, i.e.

$$D(X\|\mathrm{Po}(\lambda)) \geq D(\mathrm{Po}(\mu)\|\mathrm{Po}(\lambda)) = \mu\left(\frac{\lambda}{\mu} - 1 - \log\frac{\lambda}{\mu}\right).$$

For $\mu \leq \lambda$ we get

$$D(X\|\mathrm{Po}(\lambda)) \geq \frac{(\lambda-\mu)^2}{2\lambda} = \frac{E[P_1(X)]^2}{2}. \tag{4}$$

We conjecture that a result similar to (4) holds for any order of the Poisson-Charlier polynomial.

Conjecture 5: For any random variable X with values in $\mathbb{N}_0$ and for any $k \in \mathbb{N}$ we have

$$D(X\|\mathrm{Po}(\lambda)) \geq \frac{E[P_k(X)]^2}{2} \tag{5}$$

if $E[P_k(X)] \leq 0$.

We have not been able to prove this conjecture but we can prove the following weaker result.

Theorem 6: For any random variable X with values in $\mathbb{N}_0$ and any $k \in \mathbb{N}$ there exists $\varepsilon > 0$ such that for $E[P_k(X)] \in [-\varepsilon; 0]$ inequality (5) holds.

Proof: Define $c = E[P_k(X)]$. For c fixed the divergence $D(X\|\mathrm{Po}(\lambda))$ is minimal for the distribution Po_β given by

$$\mathrm{Po}_\beta(x) = \frac{\exp(-\beta \cdot P_k(x))}{Z(\beta)} \cdot \mathrm{Po}(\lambda, x) \ , \tag{6}$$

where Z is the partition function

$$Z(\beta) = \sum_{x=0}^{\infty} \exp(-\beta \cdot P_k(x)) \cdot \mathrm{Po}(\lambda, x)$$

and again $\beta \geq 0$ is determined by the condition $c = E[P_k(X)]$. We observe that $Z(0) = 1$ and that $Z(\beta) > 0$. From now on we shall consider c as a function of β.

We need the derivative of the partition function

$$Z'(\beta) = -\sum_{x=0}^{\infty} P_k(x) \cdot \mathrm{Po}_\beta(x) \cdot Z(\beta) = -c \cdot Z(\beta)$$

and see that $Z'(\beta) \geq 0$, and therefore $Z(\beta) \geq 1$ for all $\beta \geq 0$. We also see that $c = -Z'(\beta)/Z(\beta)$. The divergence can be written as

$$D(\mathrm{Po}_\beta\|\mathrm{Po}(\lambda)) = E_{\mathrm{Po}_\beta}\left[\log\frac{d\mathrm{Po}_\beta}{d\mathrm{Po}}\right] = -\beta c - \log(Z) \ .$$

We have to prove that $D \geq c^2/2$, which is obvious for $\beta = 0$. We take derivatives with respect to β and get

$$\frac{dD}{d\beta} = -c - \beta\frac{dc}{d\beta} - \frac{Z'}{Z} = -\beta\frac{dc}{d\beta}, \ \frac{d}{d\beta}\left(\frac{c^2}{2}\right) = c\frac{dc}{d\beta}.$$

Therefore it is sufficient to prove that $-\beta\frac{dc}{d\beta} \geq c\frac{dc}{d\beta}$, which is equivalent to $-c \leq \beta$ because $\frac{dc}{d\beta} < 0$. If $c \in [-\varepsilon; 0]$ then the theorem only has to be proved for $\beta \in [0; \varepsilon]$. Now $-c = \frac{Z'}{Z} \leq Z'$, so it is sufficient to prove that $Z'(\beta) \leq \beta$. For $\beta = 0$ it is obvious so it is sufficient to prove that $Z'' \leq 1$. For $\beta = 0$ we have $Z''(0) = E_{\mathrm{Po}(\lambda)}\left[(P_2(X))^2\right] = 1$. Therefore it is

sufficient to prove that $\frac{d^3Z}{d\beta^3} \leq 0$ for $\beta \in [0;\varepsilon]$, but if ε is chosen sufficiently small then this is true because $\frac{d^3Z}{d\beta^3}(0) = -E_{\mathrm{Po}(\lambda)}\left[(P_2(X))^3\right] < 0$ which is proved in the appendix. ∎

Conjecture 5 can be proved for $k = 2$. A short version of a much longer proof can be found in the appendix.

Theorem 7: For any random variable X with values in $\mathbb{N}_0$ then inequality (5) holds for $k = 2$ if $E[P_2(X)] \leq 0$.

V. Asymptotic lower bounds

This section combines results from Section II, III, and IV. Let κ denote the first value of k such that $E[P_k(X)] \neq 0$ and put $c = E[P_\kappa(X)]$. Lower bounds on the rate of convergence are essentially given in terms of κ and c.

First we shall see how the factorial and Poisson-Charlier moments scale under convolution and thinning. Use of Vandermonde Identity for factorials combined with equation (1) leads to the following lemma.

Lemma 8: Let $X_1, X_2, ...$ be a sequence of independent identically distributed discrete random variables all distributed like X. If $\gamma = E\left[X^{\underline{k_0}}\right]$ then

$$E\left[\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\right)^{\underline{k}}\right] = \begin{cases} \lambda^k, & \text{for } k < \kappa; \\ \lambda^k + \frac{\gamma - \lambda^k}{n^{\kappa-1}}, & \text{for } k = \kappa; \\ \lambda^{\kappa+1} + \frac{(n-1)(\kappa+1)\lambda(\gamma-\lambda^\kappa)}{2n^\kappa} + \frac{E\left[X^{\underline{k_0+1}}\right]-\lambda^{\kappa+1}}{n^\kappa}, & \text{for } k = \kappa+1. \end{cases}$$

The Poisson-Charlier moments satisfy

$$E\left[P_k\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\right)\right] = \begin{cases} 0, & \text{for } k < \kappa; \\ \frac{E[P_k(X)]}{n^{k-1}}, & \text{for } k = \kappa, \kappa+1. \end{cases}$$

We now present lower bounds on the rate of convergence in the Law of Thin Numbers in the sense of information divergence. The key idea is that we bound $D(P\|\mathrm{Po}(\lambda)) \geq D(P^*\|\mathrm{Po}(\lambda))$, where P^* is the minimum information distribution in a class containing P. Using the construction for P^* found in Section III, we often can find an explicit expression for the right hand side

Theorem 9: Let X be a random variable with values in N_0. If $E[P_\kappa(X)] \leq 0$ then

$$n^{2\kappa-2} D\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\middle\|\mathrm{Po}(\lambda)\right) \geq \frac{E[P_\kappa(X)]^2}{2}. \quad (7)$$

Proof: For $k = \kappa$ there exists $\varepsilon > 0$ such that inequality (5) holds when the condition in Theorem 6 is fulfilled. Now,

$$E\left[P_\kappa\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\right)\right] = \frac{E[P_\kappa(X)]}{n^{\kappa-1}} \in [-\varepsilon; 0]$$

for sufficiently great values of n implying that

$$D\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\middle\|\mathrm{Po}(\lambda)\right) \geq \frac{1}{2}\left(\frac{E[P_\kappa(X)]}{n^{\kappa-1}}\right)^2,$$

and asymptotic lower bound (7) follows. ∎

If the distribution of X is ultra log-concave we automatically have $E[P_\kappa(X)] \leq 0$, and we conjecture that the asymptotic lower bound is tight for ultra log-concave distributions.

A similar lower bound on rate of convergence can be achieved even if $E[P_\kappa(X)] > 0$ but then it requires the existence of a moment of higher order to "stabilize" the moment of order κ. Thus we shall assume the existence of moments of all orders less than or equal to $\kappa + 1$.

Define $t = n^{1-\kappa}$. Then

$$E\left[P_\kappa\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\right)\right] = t \cdot E[P_\kappa(X)]$$

$$E\left[P_{\kappa+1}\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\right)\right] = t^{\frac{\kappa}{\kappa-1}} \cdot E[P_{\kappa+1}(X)].$$

Let P_t^* denote the minimum information distribution satisfying

$$E_{P_t^*}[P_\kappa] = a \text{ and } E_{P_t^*}[P_{\kappa+1}] = b$$

where

$$a = t \cdot E[P_\kappa(X)] \text{ and } b = t^{\frac{\kappa}{\kappa-1}} \cdot E[P_{\kappa+1}(X)].$$

Then P_t^* is an element in the exponential family and

$$\frac{P_t^*(j)}{\mathrm{Po}(\lambda, j)} = \frac{\exp(\beta_1 P_\kappa(j) + \beta_2 P_{\kappa+1}(j))}{Z(\beta_1, \beta_2)}$$

where $Z(\beta_1, \beta_2)$ is the partition function and β_1 and β_2 are determined by the conditions. Thus

$$\begin{aligned} &D(P_t^*\|\mathrm{Po}(\lambda)) \\ &= \sum_{x=0}^{\infty} P_t^*(x) \log \frac{\exp(\beta_1 P_\kappa(x) + \beta_2 P_{\kappa+1}(x))}{Z(\beta_1, \beta_2)} \\ &= \sum_{x=0}^{\infty} (\beta_1 P_\kappa(x) + \beta_2 P_{\kappa+1}(x)) P_t^*(x) - \log Z(\beta_1, \beta_2) \\ &= \beta_1 t E[P_\kappa(X)] + \beta_2 t^{\frac{\kappa}{\kappa-1}} \cdot E[P_{\kappa+1}(X)] - \log Z(\beta_1, \beta_2). \end{aligned}$$

Therefore

$$\frac{d}{dt} D(P_t^*\|\mathrm{Po}(\lambda)) = \left(\begin{array}{c|c} \frac{da}{dt} & \frac{\partial D}{\partial a} \\ \frac{db}{dt} & \frac{\partial D}{\partial b} \end{array}\right)$$

where $(\cdot \mid \cdot)$ denotes the inner product. Thus

$$\begin{aligned} &\frac{d^2}{dt^2} D(P_t^*\|\mathrm{Po}(\lambda)) \\ &= \left(\begin{array}{c|c} t\frac{d^2a}{dt^2} & t^{-1}\frac{\partial D}{\partial a} \\ t\frac{d^2b}{dt^2} & t^{-1}\frac{\partial D}{\partial b} \end{array}\right) + \left(\begin{array}{c|cc|c} \frac{da}{dt} & \frac{\partial^2 D}{\partial a^2} & \frac{\partial^2 D}{\partial a \partial b} & \frac{da}{dt} \\ \frac{db}{dt} & \frac{\partial^2 D}{\partial b \partial a} & \frac{\partial^2 D}{\partial b^2} & \frac{db}{dt} \end{array}\right) \\ &\to E[P_\kappa(X)] \frac{\partial^2 D}{\partial a^2} E[P_\kappa(X)] = E[P_\kappa(X)]^2, \end{aligned}$$

where the physics notation $(\vec{u} \mid A \mid \vec{v})$ for $(\vec{u} \mid A\vec{v})$ is used when A is a matrix. Hence,

$$\liminf_{n\to\infty} n^{2\kappa-2} D\left(T_{1/n}\left(\sum_{j=1}^{n} X_j\right)\middle\|\operatorname{Po}(\lambda)\right) \geq \liminf t^{-2} D\left(P_t^*\|\operatorname{Po}(\lambda)\right) \geq \frac{E\left[P_\kappa(X)\right]^2}{2}.$$

VI. Discrete and continuous distributions

The α-thinning $T_\alpha(P)$ of a distribution P on $\mathbb{N}_0$ is also a distribution on $\mathbb{N}_0$. We can extend the thinning operation for distributions P of random variables Y on $\mathbb{N}_0/n = \{0, \frac{1}{n}, \frac{2}{n} \ldots\}$, by letting $T_\alpha(P)$ be the distribution of $\frac{1}{n}\sum_{j=1}^{nY} B_j$, where the B_j are as before. More generally, starting with a random variable Y with distribution P on $[0,\infty)$, let P_n denote the uniformly quantized version of P supported on $\mathbb{N}_0$. It is easy to see that $T_\alpha(P_n)$ converges to the distribution of αY as $n \to \infty$. In this sense, thinning can be interpreted as a discrete analog of the scaling operation for continuous random variables.

Let $\Phi\left(\mu, \sigma^2\right)$ denote the distribution of a Gaussian random variable with mean μ and variance σ^2. We are interested in a lower bound on $D\left(X\|\Phi\left(\mu,\sigma^2\right)\right)$ in terms of the variance of X, where X is some random variable. We shall assume that $\operatorname{Var}(X) \leq \sigma^2$. First we remark that

$$D\left(X\|\Phi\left(\mu,\sigma^2\right)\right) = D\left(aX+b\|\Phi\left(a\mu+b, a^2\sigma^2\right)\right)$$

for real constants a and b. The constants a and b can be chosen so that $a\mu + b = a^2\sigma^2$. Our next step is to discretize

$$D\left(aX+b\|\Phi\left(a\mu+b,a^2\sigma^2\right)\right) \approx D\left(\lfloor aX+b\rfloor\|\operatorname{Po}\left(a^2\sigma^2\right)\right).$$

Next we use Theorem 7 to get

$$D\left(\lfloor aX+b\rfloor\|\operatorname{Po}\left(a^2\sigma^2\right)\right) \geq \frac{E\left[P_2\left(\lfloor aX+b\rfloor\right)\right]^2}{2} = \frac{1}{4}\left(\frac{\operatorname{Var}\lfloor aX+b\rfloor}{a^2\sigma^2}-1\right)^2 = \frac{1}{4}\left(\frac{\operatorname{Var}\left(\frac{\lfloor aX+b\rfloor}{a}\right)}{\sigma^2}-1\right)^2.$$

Finally we use that $\operatorname{Var}\left(\lfloor aX+b\rfloor/a\right) \to \operatorname{Var}(X)$ for $n\to\infty$ to get

$$D\left(X\|\Phi\left(\mu,\sigma^2\right)\right) \geq \frac{\left(\frac{\operatorname{Var}(X)}{\sigma^2}-1\right)^2}{4} = \frac{E\left[H_2\left(X-E[X]\right)\right]^2}{2},$$

where H_2 is the second Hermite polynomial. This inequality can also be proved by a straightforward calculation in the exponential family of Gaussian distributions. Following the same kind of reasoning we get the following new and non-trivial result:

Theorem 10: For any random variable X with mean 0 and variance 1 and for any $l \in \mathbb{N}$ there exists $\varepsilon > 0$ such

$$D\left(X\|\Phi\left(\mu,\sigma^2\right)\right) \geq \frac{E\left[H_{2l}(X)\right]^2}{2}$$

if $E\left[H_{2l}(X)\right] \in [-\varepsilon; 0]$.

If Conjecture 5 holds then the condition $E\left[H_{2l}(X)\right] \in [-\varepsilon;0]$ in Theorem 10 can be replaced by the condition $E\left[H_{2l}(X)\right] \leq 0$. The case $l=1$ has been discussed above and the case $k=2$ has also been proved [10, Thm. 7].

References

[1] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*. Oxford Studies in Probability 2, Oxford: Clarendon Press, 1992.

[2] P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2039–2041, July 2001.

[3] I. Kontoyiannis, P. Harremoës, and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. Inform. Theory*, vol. 51, pp. 466–472, Feb. 2005.

[4] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[5] A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1491–1498, June 2003.

[6] P. Harremoës and P. Ruzankin, "Rate of convergence to Poisson law in terms of information divergence," *IEEE Trans. Inform Theory*, vol. 50, pp. 2145–2149, Sept. 2004.

[7] O. Johnson, M. Madiman, and I. Kontoyiannis, "Fisher information, compound Poisson approximation, and the Poisson channel," in *Proceedings International Symposium on Information Theory 2007, Nice*, pp. 976–980, June 2007.

[8] P. Harremoës, O. Johnson, and I. Kontoyiannis, "Thinning and the law of small numbers," in *Proceedings International Symposium on Information Theory 2007, Nice*, pp. 1491–1495, IEEE Information Theory Society, June 2007.

[9] O. Johnson, "Log-concavity and the maximum entropy property of the Poisson distribution," *Stochastic Processes and their Applications*, vol. 117, no. 6, pp. 791–802, 2007.

[10] P. Harremoës, "Lower bounds for divergence in the central limit theorem," in *General Theory of Information Transfer and Combinatorics* (R. Ahlswede, L. Bäumer, N. Cai, H. K. Aydinian, V. Blinovsky, C. Deppe, and H. Mashurian, eds.), vol. 4123 of *Lecture Notes in Computer Science*, pp. 578–594, Berlin Heidelberg: Springer-Verlag, 2006.

[11] V. I. Khokhlov, "Polynomials orthogonal with respect to the multinomial distribution and the factorial-power formalism," *Theory Probab. Appl.*, vol. 46, no. 3, pp. 529–536, 2002.

VII. Appendix: technical lemmas and proofs

Proof of Lemma 3: Let $\vec{G} \in \mathbb{R}^{l-1}$ be a vector and let $C_{\vec{G}}$ be the set of distributions satisfying the following inequalities

$$E\left[P_l(X) - h_l - \sum_{k<l} G_k \cdot \left(P_k(X) - h_k\right)\right] \leq 0.$$

We see that the set $C_{\vec{G}}$ is closed because $P_l(x) - h_l - \sum_{k<l} G_k\left(P_k(x)-h_k\right) \to \infty$ for $x \to \infty$. Therefore the intersection $K = \bigcap_{\vec{G}\in\mathbf{R}^{l-1}} C_{\vec{G}}$ is closed. There exists a distribution $P^* \in K$ such that the information divergence $D\left(P\|\operatorname{Po}(\lambda)\right)$ is minimal because K is closed. ■

Proof of Theorem 4:

Case 1. If a minimizer existed it would be an element of the corresponding exponential family, but the partition function cannot be finite because $h_l > 0$ and $l \geq 2$.

For cases 2 and 3 let $P = P^*$ be the minimum information distribution satisfying the conditions (2).

Case 2. Assume that $h_k = 0$ for $k < l$ and $h_l < 0$. Assume also that $E_{P^*}\left[P_l(X)\right] < h_l$. Define $P^\theta = \theta P^* +$

$(1-\theta)\operatorname{Po}(\lambda)$. Then the conditions (2a) holds for $P = P^\theta$ and

$$E_{P^\theta}[P_l(X)] = \theta E_{P^*}[P_l(X)] + (1-\theta) E_{\operatorname{Po}(\lambda)}[P_l(X)] = \theta E_{P^*}[P_l(X)].$$

Thus $E_{P^\theta}[P_l(X)] = h_l$ if $\theta = \frac{h_l}{E_{P^*}[P_l(X)]} \in]0,1[$. Therefore P^θ satisfies (3) but $D\left(P^\theta \| \operatorname{Po}(\lambda)\right) \leq \theta D\left(P^* \| \operatorname{Po}(\lambda)\right) < D\left(P^* \| \operatorname{Po}(\lambda)\right)$ and we have a contradiction.

Case 3. Now, assume that $h_k = 0$ for $k < l-1$ and $h_{l-1} > 0$. Moreover, assume that $E_{P^*}[P_l(X)] < h_l$. Using the result of case 1 we see that there exists a distribution $\tilde{P}$ for which the l first moments exist and that the first $l-1$ moments satisfy (2a) but with $D\left(\tilde{P} \| \operatorname{Po}(\lambda)\right) < D\left(P^* \| \operatorname{Po}(\lambda)\right)$. Define $P^\theta = \theta P^* + (1-\theta)\tilde{P}(\lambda)$. Then the conditions (2a) holds for $P = P^\theta$ and

$$E_{P^\theta}[P_l(X)] = \theta E_{P^*}[P_l(X)] + (1-\theta)\theta E_{\tilde{P}}[P_l(X)].$$

Therefore $E_{P^\theta}[P_l(X)] \leq h_l$ for θ sufficiently close to 1 but $D\left(P^\theta \| \operatorname{Po}(\lambda)\right) \leq \theta D\left(P^* \| \operatorname{Po}(\lambda)\right) < D\left(P^* \| \operatorname{Po}(\lambda)\right)$ and we have a contradiction. Therefore P^* satisfies $E[P_l(X)] = h_l$. ∎

Proof of Theorem 7: First we note that $P_2(x) \geq -2^{-1/2}$ for any $x \in \mathbb{N}_0$. Hence, $E[P_2(X)] \in \left[-2^{-1/2}; 0\right]$. Then, according to the proof of Theorem 6 it is sufficient to prove that $\frac{d^3 Z}{d\beta^3} \leq 0$ for $\beta \in \left[0; 2^{-1/2}\right]$. The function $\beta \curvearrowright \frac{d^2 Z}{d\beta^2} = \sum_{x=0}^{\infty} P_2(x)^2 \exp\left(-\beta P_2(x)\right) \operatorname{Po}(\lambda, x)$ is convex, so it is sufficient to prove the inequality $\frac{d^2 Z}{d\beta^2} \leq 1$ for a single value $\beta = \beta_0 \geq 2^{-1/2}$. Consider the function $f(x) = x^2 \exp(-\beta x)$ with $f'(x) = (2-\beta x)\, x \exp(-\beta x)$.The function f is decreasing for $x \leq 0$, has minimum for $x = 0$, increases for $0 \leq x \leq 2/\beta$, has a local maximum $4\exp(-2)/\beta^2$ in $x = 2/\beta$ and decreases for $x \geq 2/\beta$. We solve the equation $\frac{4\exp(-2)}{\beta_0^2} = 1$ and obtain $\beta_0 = 2/e = 0.735\,76\cdots$ and note that $\beta_0 \geq 2^{-1/2}$. We see that it is sufficient to prove that $\sum_{x=0}^{\infty} P_2(x)^2 \exp\left(-\beta_0 P_2(x)\right) \operatorname{Po}(\lambda, x) \leq 1$.

The graph of $x \to P_2(x)$ is a parabola and we have $P_2(\lambda) = P_2(\lambda+1) = -2^{-1/2}$. For all values $x \notin]\lambda; \lambda+1[$ we have $f(h(x)) \leq \max\left\{1, f\left(-2^{-1/2}\right)\right\} = 1$ and x can only assume integer values so at most one value of x will contribute to the mean value with a value greater that 1. A careful inspection of different cases will show that the single value of x that may contribute to the mean with greater than 1 will be averaged out with some other value of x. ∎

Lemma 11 (Khoklov [11]): [1]

$$P_k(x) P_l(x) = (-1)^{k+l} \left(\frac{\lambda^{k+l}}{k!l!}\right)^{1/2} \sum_{m=0}^{k+l} c_m P_m(x)$$

where c_m as a function of k, l and λ is given by

$$\sum_{n=0}^{m} \frac{\sum_{\mu=0}^{k}\sum_{\nu=0}^{l} \binom{k}{\mu}\binom{l}{\nu} \mu^{\underline{n}} \nu^{\underline{n}} (\mu+\nu-n)^{\underline{m}} (-1)^{\mu+\nu}}{n!\lambda^n \left(m!\lambda^m\right)^{1/2}}.$$

[1] The original formula by Khokhlov [11] contains an error in that the factor $(-1)^{k+l}$ is missing but his proof is correct.

Lemma 12: For a Poisson random variable X with mean value λ we have $E\left[P_k(X)^3\right] > 0$ for any $k \in \mathbb{N}$.

Proof: According to Lemma 11 we have

$$E\left[\left(P_k(X)\right)^3\right] = E\left[\left((-1)^{k+k} \frac{\lambda^k}{k!} \sum_{m=0}^{k+k} c_m P_m(X)\right) P_k(X)\right] = \frac{\lambda^k}{k!} c_k$$

where c_k is defined in Lemma 11. Therefore it is sufficient to prove that

$$\sum_{\mu=0}^{k}\sum_{\nu=0}^{k} \binom{k}{\mu}\binom{k}{\nu} \mu^{\underline{n}} \nu^{\underline{n}} (\mu+\nu-n)^{\underline{k}} (-1)^{\mu+\nu} \geq 0$$

for all $0 \leq n \leq k$ and that there exists at least one value of n such that the left hand side is positive. For any n satisfying $0 \leq n \leq k$ we can use the Vandermonde Identity for factorials to get

$$\sum_{\mu=0}^{k}\sum_{\nu=0}^{k} \binom{k}{\mu}\binom{k}{\nu} \mu^{\underline{n}} \nu^{\underline{n}} (\mu+\nu-n)^{\underline{k}} (-1)^{\mu+\nu} = \sum_{\substack{a+b+c \\ =k}} \begin{pmatrix} & k & \\ a & b & c \end{pmatrix} n^{\underline{c}} \begin{pmatrix} \sum_{\mu=0}^{k} \binom{k}{\mu} \mu^{\underline{n}} (\mu-n)^{\underline{a}} (-1)^{\mu} \\ \times \\ \sum_{\nu=0}^{k} \binom{k}{\nu} \nu^{\underline{n}} (\nu-n)^{\underline{b}} (-1)^{\nu} \end{pmatrix}.$$

Now

$$\begin{aligned} &\sum_{\mu=0}^{k} \binom{k}{\mu} \mu^{\underline{n}} (\mu-n)^{\underline{a}} (-1)^{\mu} \\ &= \sum_{\mu=n+a}^{k} \frac{k^{\underline{n+a}} (k-n-a)!}{(\mu-n-a)!\,(k-\mu)!} (-1)^{\mu} \\ &= k^{\underline{n+a}} (-1)^{n+a} \sum_{\mu=n+a}^{k} \binom{k-n-a}{\mu-n-a} (-1)^{\mu-n-a} \\ &= \begin{cases} 0, & \text{for } a \neq k-n; \\ k!\,(-1)^k, & \text{for } a = k-n. \end{cases} \end{aligned}$$

This sum is only non-zero if $a = k-n$. Similarly the sum $\sum_{\nu=0}^{k} \binom{k}{\nu} \nu^{\underline{n}} (-1)^{\nu} (\nu-n)^{\underline{b}}$ is only non-zero if $b = k-n$. Thus $c = k - 2(k-n) = 2n-k$ and the condition $c \geq 0$ implies that $n \geq k/2$. Hence

$$\begin{aligned} &\sum_{\mu=0}^{k}\sum_{\nu=0}^{k} \binom{k}{\mu}\binom{k}{\nu} \mu^{\underline{n}} \nu^{\underline{n}} (\mu+\nu-n)^{\underline{k}} (-1)^{\mu+\nu} \\ &= \begin{pmatrix} & k & \\ k-n & k-n & 2n-k \end{pmatrix} n^{\underline{2n-k}} k!\,(-1)^k\, k!\,(-1)^k \\ &= \left(k^{\underline{n}}\right)^3 n^{\underline{k-n}}, \end{aligned}$$

which is always non-negative and it is positive if $k/2 \leq n \leq k$. ∎